Audio Engineering Society
Convention Paper
Presented at the 112th Convention
2002 May 10–13 Munich, Germany

Why Binaural Cue Coding is better than Intensity Stereo Coding

Frank Baumgarte[1], Christof Faller[1]

[1] *Media Signal Processing Research, Agere Systems, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.*

Correspondence should be addressed to Frank Baumgarte (`fb@agere.com`)

ABSTRACT

Intensity Stereo Coding (ISC) is a joint-channel audio coding tool that is part of the ISO/MPEG standards. ISC can introduce severe distortions if applied to the full audio bandwidth or to audio signals with a dynamic or wide spatial image. In contrast, Binaural Cue Coding (BCC) is a systematic approach for representing auditory spatial cues which includes ISC as a subset. BCC is independent of the time/frequency resolution used by the coder, thus it can be optimized for spatial image reproduction. Subjective listening tests confirm that ISC is significantly compromised by an inappropriate time/frequency resolution and that BCC has superior quality and robustness.

1 INTRODUCTION

Intensity Stereo Coding (ISC) is a joint-channel coding tool that is part of the ISO/MPEG family of standards [1, 2]. ISC is applied to reduce irrelevant information of audio channel pairs and originated from [3]. In each coder sub-band that uses ISC, the sub-band signals are replaced by a sum signal and a direction angle (azimuth). The azimuth controls the intensity stereo position of the phantom source created at the decoder. Only one azimuth is transmitted for a scalefactor band (1024 coder bands are divided into roughly 50 scalefactor bands that are spaced proportional to auditory critical bands). ISC is capable of significantly reducing the bit rate for stereo and multi-channel audio where it is used for channel pairs. However, its application is limited since intolerable distortions can occur if ISC is used for the full bandwidth or for audio signals with a highly dynamic

and wide spatial image [4]. Potential improvements of ISC are constrained since the time-frequency resolution is given by the core audio coder and cannot be modified without adding considerable complexity.

Binaural Cue Coding (BCC) is a systematic approach to represent auditory spatial cues contained in a stereo or multi-channel audio signal. For the rationale of BCC as well as alternative implementations of BCC and applications the reader is referred to [5, 6, 7]. BCC can be applied to audio coding as a preprocessor that represents the stereo or multi-channel signal by a mono signal plus encoded spatial cues. Figure 1 shows the BCC analyzer which generates BCC parameters and the summation used to generate a mono signal. The mono signal is subsequently compressed by any suitable audio coder. At the decoder site, the BCC parameters allow to reconstruct the spatial image by inserting spatial cues while generating the stereo signal. As opposed to the coder-integrated ISC, BCC is in general not constrained in terms of the time-frequency resolution used for the audio compression. Therefore, all parameters of the BCC scheme can be optimized independently from the subsequent mono audio compression scheme.

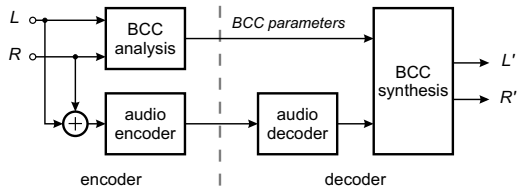

Fig. 1: BCC-based stereo audio coding scheme.

ISC exploits only interaural or inter-channel level-difference cues for the reconstruction of spatial images. BCC supports not only these level-difference cues, but also time-difference cues or head-related transfer functions [8] to reconstruct auditory spatial images [7]. Thus, ISC can be interpreted as a special case of BCC with the restriction of using only level-difference cues and using the given time/frequency resolution of the audio coder.

This paper investigates the type and amount of potential audio quality degradation introduced by ISC and BCC. The experiments reported include an ISC configuration equivalent to MPEG-2 AAC [2] and BCC with different time/frequency resolutions. The degradations are assessed in a subjective listening test. A theoretical analysis highlights the drawbacks of the integration of ISC into the audio compression scheme. This includes an analysis of the amount of aliasing distortions created by ISC and BCC.

2 COMPARISON OF INTENSITY STEREO AND BINAURAL CUE CODING

For a meaningful performance comparison between ISC and BCC, BCC is restricted to employ only level differences as auditory spatial cues. For loudspeaker playback this restriction has only minor impact since in this case level differences are the most important cues. Therefore, the optional use of time-difference cues or head-related transfer functions in BCC will not be discussed further.

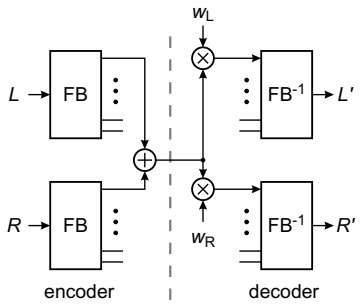

Fig. 2: Generic block diagram of Intensity Stereo processing in the sub-band domain using analysis filter bank (FB) and synthesis filter bank (FB^{-1}).

Figure 2 shows an Intensity Stereo processing scheme of a sub-band signal at the encoder and decoder site ignoring quantization and coding. This generic scheme applies to both ISC and BCC. For BCC this becomes obvious when we move the summation operation of L and R in Fig. 1 into the audio encoder block, such that the summation is applied in each sub-band. The BCC parameters include the necessary data to derive the weighting factors w_{L} and w_{R}.

At the encoder of Fig. 2, the sum of the left and right channel is computed and transmitted. Additionally, the azimuth is provided, so that the decoder can derive the weighting factors w_{L} and w_{R} for the two audio channels to control the phantom image position. For ISC the MPEG–AAC Standard [2] recommends to adjust the weights such that the energy of the sum signal is equal to the summed energies of the decoded signal channels:

$$w_{\mathrm{L},m} = \sqrt{\frac{E_{\mathrm{L},m}}{E_{\mathrm{L},m} + E_{\mathrm{R},m}}} \tag{1}$$

$$w_{\mathrm{R},m} = \sqrt{\frac{E_{\mathrm{R},m}}{E_{\mathrm{L},m} + E_{\mathrm{R},m}}}. \tag{2}$$

The energies of the left and right channel in sub-band m at the encoder are denoted $E_{\mathrm{L},m}$ and $E_{\mathrm{R},m}$ respectively.

BCC uses the same energy conservation formula. However, the different time-frequency resolution of ISC and BCC is crucial for achieving sufficient audio quality. ISC is integrated in the core audio coder, thus needs to work with the given time/frequency resolution of the modified discrete cosine transform (MDCT) of MPEG–AAC or the hybrid filter bank of MPEG-1 Layer 3 [1]. In contrast, BCC is based on an FFT that can be adjusted in size and overlap to specific applications.

2.1 Properties of MDCT and FFT

Many recent sub-band audio coders use the MDCT because it provides a critically sampled sub-band representation of the time-domain signal based on overlapping time-domain windows. Critical sampling is a key to get a good compression performance. The sub-band signals of the MDCT are down-sampled by a factor equal to the number of bands with respect to the input sampling rate. Overlapping windows smooth out misalignments of the decoder-reconstructed frames. MPEG-AAC specifies two basic time-domain window shapes that can be used alternatively. For each frame the same window must be used in the encoder and decoder. In the following, only the sinusoidal window shape is considered.

For long blocks, MPEG-AAC has a window length of 2048 samples and a window overlap of 1024 time-domain samples. The MDCT represents each block by 1024 sub-band samples (one sample in each of the 1024 bands). For short blocks, these numbers are divided by 8. Figure 3 shows the magnitude frequency responses of an MDCT sub-band together with the aliasing components created by the neighboring bands. Each sub-plot shows bands 14, 15, and 16 of a 128-band MDCT. The total number of 128 MDCT bands corresponds to MPEG-AAC in short block mode that uses a time-window length of 256. The frequency responses are approximately invariant with respect to a shift of band numbers. Therefore, Fig. 3 also applies to band triples different from that centered at band 15. Figure 3A outlines the aliasing components for critical sampling (window overlap of 128 samples). Since ISC can substantially modify the sub-band levels of neighboring critical bands, it potentially introduces aliasing distortions in the vicinity of scalefactor-band boundaries.

Aliasing can be reduced by increasing the sub-band sampling rate, i.e. to decrease the effective sub-band down-sampling factor. This is done by increasing the window overlap which results in less-than-critical sampling. Figure 3B shows that the aliasing components for 2-times oversampling (window overlap of 192) are reduced with respect to critical sampling. The resulting sub-band sampling rate is denoted f_S. The dashed and dashed-dotted components within the indicated bandwidth of $f_S/2$ outline the level of frequency-domain aliasing introduced by an inverse MDCT. The amount of aliasing shown refers to a situation where the input signal of the analysis MDCT has constant spectral energies throughout the three neighboring bands and the sub-band samples of the two outer bands are set to zero before applying the inverse MDCT. For other situations the amount of aliasing depends on the spectral shape and the spectral modifications applied. Figure 3C shows that the aliasing components are further reduced by an increased oversampling factor of 4. For a window length of 2048, the frequency responses and aliasing shown in Fig. 3 are virtually identical to the responses shown for a length of 256.

For an efficient synthesis of the time-domain stereo signal, the BCC decoder is based on an inverse FFT as presented in [5]. In contrast to the MDCT, the spectral or sub-band representation derived from an FFT is only critically sampled if there is no window overlap. In this case a rectangular window is appropriate for analysis and synthesis. If the signal is modified in the sub-bands, the inverse FFT synthesis will create "blocking" artifacts due to the abrupt block boundaries without overlap. Applying the window sequence of the MDCT increases the sub-band sampling rate by a factor of 2. The corresponding window sequence for the analysis FFT and synthesis inverse FFT is shown in Fig. 4A. This configuration avoids blocking artifacts and has virtually identical aliasing as the sub-band aliasing situation shown in Fig. 3B. For comparison, the window sequence for an MDCT with oversampling of 2 is shown in Fig. 4C. Obviously, the MDCT uses only half the window shift of the FFT. Thus, it appears as having a higher time resolution. This is outlined in Figure 4B and 4D which indicate each successive block of sub-band samples at the time instance of the center of the corresponding time-domain window. In this configuration the FFT represents M time-domain samples by $2M$ sub-band samples while the MDCT represent them by two blocks of M sub-band samples. Thus, there are two samples per sub-band. But only the MDCT implements the proper time shift of $M/2$ between the two samples.

A relevant difference between the FFT and MDCT for Intensity Stereo processing results from the reconstruction properties of both transforms. The inverse FFT uses the real and imaginary part of the spectrum to reconstruct the time-domain signal. Spectral modifications that only change the magnitudes will not affect the reconstructed signal phase. Therefore, the overlap-add of two subsequent blocks is perfectly phase aligned. In contrast, the inverse MDCT is based on time-domain aliasing cancellation. Aliasing components in the cur-

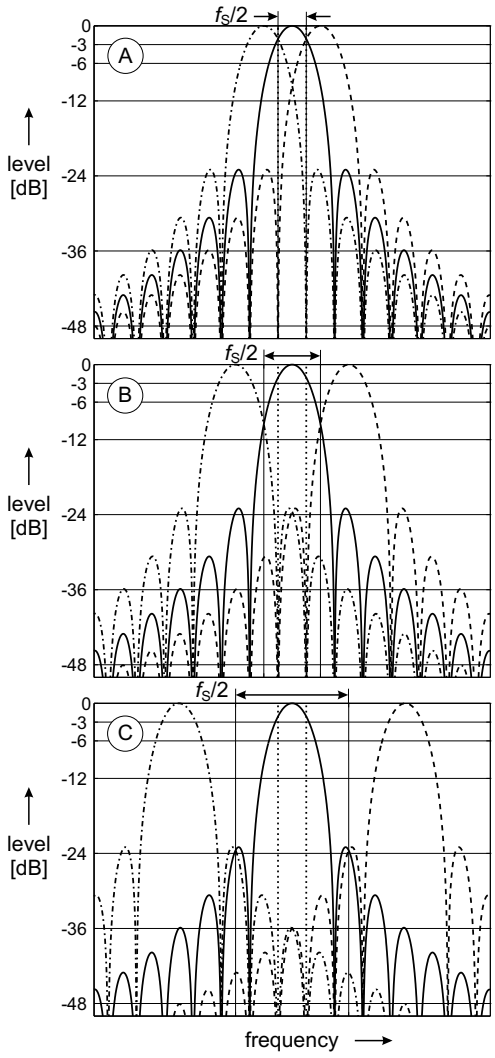

Fig. 3: Sub-band magnitude frequency response of band no. 15 (solid) of a 128-band MDCT. Aliasing components from neighboring bands (dashed, dashed-dotted). The sub-band sampling frequency is denoted f_S. **A** critical sampling, **B** oversampling by a factor of 2, **C** oversampling by a factor of 4. The dotted lines mark the borders of the transition band.

rent block are canceled by the overlapping components from the previous and next block. Modifications to the MDCT sub-band signals will thus result in time-domain aliasing. As an analog to the FFT, the MDCTs of two subsequent blocks can be interpreted alternatively as the real and imaginary spectral representation because the MDCT impulse responses are 90° out of phase. This interpretation shows that spectral modifications affect the phase alignment of overlapping blocks. Thus, a pure intensity modification can only be approximated by limiting the permitted amount of level change of the spectral weighting factors for subsequent blocks. Accordingly, the intensity stereo phantom image position can only be moved with a certain maximum speed to avoid the phase distortions.

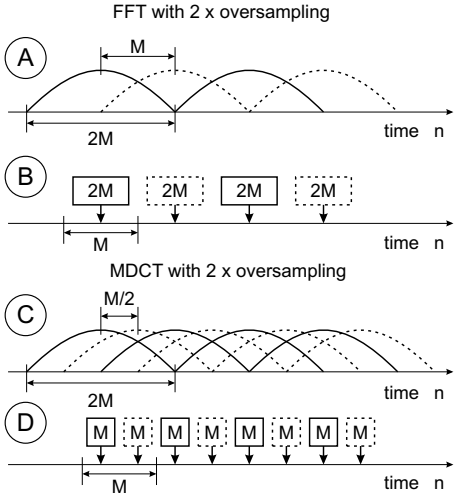

Fig. 4: Time-domain window sequence applied for FFT (**A**) and MDCT (**C**) to achieve a sub-band oversampling factor of 2. Figure **B** and **D** indicate the centers of the time intervals represented by the corresponding block of sub-band samples. The number of sub-bands is M.

Furthermore, a complexity analysis of an MDCT and FFT-based intensity stereo processing is instructive. According to [9] the computational complexity of an M-band MDCT and M-band FFT (FFT length $2M$) and 2 times oversampling can be calculated as follows:

$$C_{\mathrm{MDCT}}(\mathrm{mul}) = 2M(\log_2 M + 5) \quad (3)$$

$$C_{\mathrm{MDCT}}(\mathrm{add}) = 2M(3\log_2 M + 3) \quad (4)$$

$$C_{\mathrm{FFT}}(\mathrm{mul}) = 2M(\log_2 M - 2) + 4 \quad (5)$$

$$C_{\mathrm{FFT}}(\mathrm{add}) = 2M(3\log_2 M - 2) + 8\ . \quad (6)$$

The number of real multiplications and additions is denoted $C(\mathrm{mul})$ and $C(\mathrm{add})$, respectively. Table 1 summarizes the complexities for two window lengths corresponding to MPEG-AAC in long and short block mode. Assuming an oversampling factor of 2 for each sub-band, the FFT turns out significantly more efficient.

3 SUBJECTIVE EVALUATION

The different properties of the MDCT and FFT revealed in the previous section provide some insight why both spectral representations might result in different

perceptual quality of the synthesized stereo signal. In this section we assess different quality dimensions of the level-difference-cue-based synthesis schemes with the time/frequency resolution as parameter.

Intensity Stereo Coding (ISC) is motivated by spatial perception that chiefly relies on interaural level differences at frequencies above 1.6 kHz. At lower frequencies interaural time differences are the most important spatial cues [8]. In that respect it is interesting to note that a level difference at low frequencies applied to a loudspeaker pair appears chiefly as an interaural time difference at the listeners ears [8]. The level difference translates into a time difference due to the head size being relatively small compared to the acoustical wavelength. For that reason, we believe that inter-channel level-difference cues can reproduce a realistic spatial image even when they are applied at low frequencies of an audio signal for loudspeaker playback. This assumption is supported by the fact that most mixing consoles apply level differences to the full-band time signals of the two stereo channels to place phantom sources.

Since recordings usually contain dynamic spatial images, the stereo synthesis must be able to reproduce time-varying phantom source positions with sufficient time resolution. However, if we want to avoid the complexity of a multi-resolution representation, e.g. [10], the time resolution determines the maximum spectral resolution. Thus, it is crucial to find the critical time/frequency resolution that achieves optimum perceptual quality.

Psychoacoustics suggests that spatial perception is most likely based on a critical band representation of the acoustic input signal [8]. Since critical bandwidths of auditory filters can be approximated by a logarithmic function of frequency (constant Q) over a large frequency range, there is a mismatch between uniform filter banks (or transforms) and auditory filters. However, for complexity reasons it is desirable to use a uniform filter bank, like an MDCT or FFT. Accepting this constraint, only the time resolution or the number of frequency bands can be tuned to achieve maximum audio quality. To increase the compression efficiency of such a scheme, the limited critical band resolution of the auditory system can be exploited, for instance, by applying the weighting factors that implement the level-difference cues to scalefactor bands as in MPEG-AAC.

FB	*length*	C(mul)		C(add)	
MDCT	2048	65536		147456	
FFT	2048	36868	*56%*	126984	*86%*
MDCT	256	3072		6144	
FFT	256	1284	*42%*	4872	*79%*

Table 1: Comparison of computational complexity of MDCT and FFT with 2 times oversampling in each sub-band. The percentages indicate the ratio $C_{\mathrm{FFT}}/C_{\mathrm{MDCT}}$.

A psychoacoustic test was designed to assess the impact of different time/frequency resolutions and filter bank types in the decoder on the audio quality. The level-difference-cue parameters are not quantized and they are applied individually in each sub-band to exclude possible degradations caused by quantization or coarse frequency resolution. The only exception is one decoder which uses only one parameter per scalefactor band. The parameters are derived with an analysis scheme that uses a Cochlear Filter Bank and an estimation of inter-channel level differences. For details of the analysis scheme, the reader is referred to [6]. This analysis scheme was chosen since it represents a reasonable first approach to achieve a perceptually appropriate time/frequency resolution for the level-difference estimation. However, the derived critical-band representation needs to be mapped to the different time/frequency resolutions of the specific decoder. The mapping is done by proper resampling of the parameters.

For simplicity, the decoders were implemented as shown in Fig. 5. As opposed to Fig. 2, the mono signal is generated in the time domain. The sub-band representation of the mono signal is derived by applying an analysis filter bank that fits the synthesis filter banks at the decoder output. This scheme resembles the one in Fig. 2, however, here we use a separate estimation of level differences based on the Cochlear Filter Bank to derive the weighting factors w_L and w_R, which is not shown.

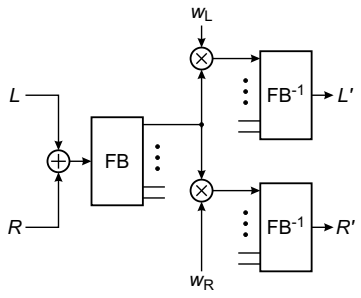

Fig. 5: Synthesis scheme used for the subjective test.

Table 2 lists the 8 decoder configurations used in the test. The filter bank (FB) type is either an MDCT or FFT. For the MDCT the time-window length was chosen according to the long and short block mode of MPEG-AAC. The long MDCT (A) uses one level cue parameter per scalefactor band. It was included in this experiment to assess the potential degradation due to the coarser fre-

quency resolution in comparison to (B). A short MDCT with 2-times oversampling (D) provides the same frequency resolution as (C) but it has reduced aliasing. All FFT decoders (E-H) have an oversampling factor of 2. Only the time/frequency resolution is varied by a factor of two for subsequent decoders.

ref.	*FB*	*length*	*oversampling factor*	*resolution*
A	MDCT	2048	1	SCF bands
B	MDCT	2048	1	full
C	MDCT	256	1	full
D	MDCT	256	2	full
E	FFT	2048	2	full
F	FFT	1024	2	full
G	FFT	512	2	full
H	FFT	256	2	full

Table 2: Decoder parameters used in the subjective test. The resolution of decoder A is given by the scalefactor (SCF) bandwidths.

Four different stereo audio excerpts, each with a duration of approximately 10 s sampled at 32 kHz, are used in the test. The excerpts are first analyzed according to [6] to estimate the inter-channel level differences. These level differences are then resampled to match the decoder time/frequency resolution and provided to the decoder along with the (mono) sum signal. The decoder synthesizes the level difference in each band when generating the reconstructed stereo signal according to Fig. 5.

The first three excerpts are mixed, each from two mono sources by applying a level difference of +10 and −10 dB, respectively. Both mono sources for a specific mix have the same category as listed in Tab. 3. The imposed fixed level difference ensures a stable and focused spatial image. (The same three excerpts were used in another assessment described in [6] were they are denoted D2, E2, and G2). The audio excerpts are selected from a collection of critical stereo signals with the objective of having different types of content and the most critical material for level cue imaging in the test. The same category was chosen for each of the first 3 excerpts because it is supposed to be more difficult to separate two similar sources and create a stable image with level-difference cues. The fourth excerpt is a stereo recording of applause. It is known as a critical signal for ISC since the spatial image is very dynamic.

The excerpts are presented over loudspeakers. The test is performed by each subject sitting at the standard listening position for conventional stereo playback. The excerpts were played back from a computer under the subject's control and with comfortable volume. The five

excerpt	*category*	*left*	*right*
1	speech	male	female
2	singing	tenor	soprano
3	percussions	castanets	drums
4	applause	(stereo recording)	

Table 3: List of the audio excerpts used in the subjective test. The last two columns contain the sources of 1, 2, and 3, that are placed to the left or right side of the spatial image by imposing a level difference (amplitude panning).

participating subjects were asked to grade different specific distortions and the overall audio quality of the processed excerpts with respect to the known reference, the unprocessed excerpt. The four different grading tasks are summarized in Tab. 4. Task 1 and 2 assess the two properties of the reproduced spatial image that are thought to determine the spatial image quality of ISC or BCC since they typically dominate all other types of spatial image distortions. Task 3 evaluates distortions introduced by the stereo synthesis that do not result in image artifacts. For example, aliasing and blocking artifacts should be detected here. Task 4 is an important measure for a global optimization of BCC.

	task	*scale*
1	image width	stereo...mono
2	image stability	stable...unstable
3	audio quality disregarding spatial image distortions	ITU-R 5-grade impairment
4	overall audio quality	ITU-R 5-grade impairment

Table 4: Tasks of the subjective test.

During the test, each subject is able to randomly access one excerpt processed by the 8 different decoders and the reference by using the corresponding "Play" button of the graphical interface shown in Fig. 6. This Play function stops an active audio output at any time, so that the subject has an opportunity to do a quick listening before it continues with a more thorough evaluation. All the slider positions can be modified at any time to reflect the proper grading and ranking of the excerpts. It is important to note that subjects were specifically asked to pay attention to the rank order of the excerpts. The feature of being able to play the excerpts according to their rank order greatly facilitates this task as opposed to other schemes that allow to listen only once to each item in a pre-defined order. For the four different test tasks, only the title and the grading scale was adapted

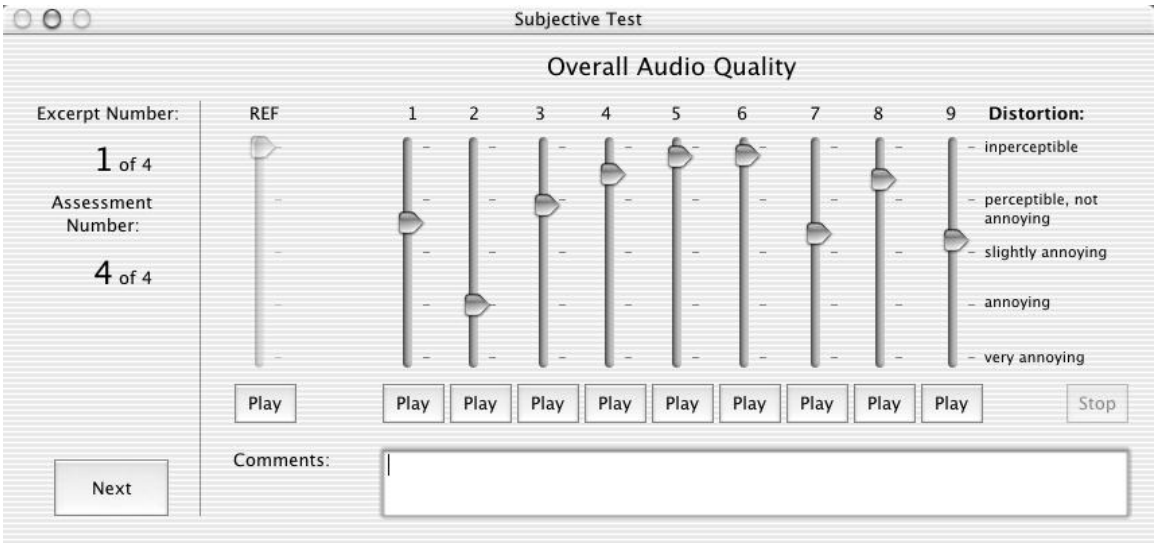

Fig. 6: Graphical user interface of subjective test, task 4.

in the graphical interface. The ordering of the decoders was randomly chosen for each subject and each excerpt, but it was not changed during the four different tasks performed on one excerpt.

4 RESULTS OF SUBJECTIVE EVALUATION

The experimental results are shown for the individual excerpts only. Averaging over the gradings from different excerpts cannot be justified due to substantially deviating ratings. The gradings of each task will be discussed in the following sub-sections.

4.1 Image Width

The gradings for image width are shown in Fig. 7 for each excerpt over the different decoders with respect to the reference.

Apparently, all coders reduce the image width for all excerpts. The largest image width reduction appears for the two 256-point MDCTs for all excerpts, except for applause were the 256-point MDCT without oversampling has relatively good performance. The low performance might result from the time-domain aliasing property of the MDCT discussed above. However, this intuitive explanation is still unproven and needs further evaluation.

The 2048-point MDCT with scalefactor-band resolution appears to have virtually the same gradings as the MDCT with maximum resolution. Both 2048-point MDCTs have similar gradings as the 2048-point FFT.

For excerpt 2 there is a trend toward a smaller image width with reduced FFT size. This trend is reverse for excerpt 3 and 4. This result can be explained by the more stationary character of excerpt 2 (singing) in contrast to the non-stationary excerpt 3 (percussions) and 4 (applause) which require a higher time resolution for a proper image reproduction. The overall performance of the 256-point FFT is significantly better compared with the MDCT of the same length.

4.2 Image Stability

Gradings for image stability are given in Fig. 8. The image stability is maximum if the virtual sound source location is stationary. For excerpts 1, 2, and 3 that location is well defined. However, for excerpt 4 (applause) each source is only active for a short time so that a moving source cannot be detected. That is why excerpt 4 appears "stable" for all coders.

From the remaining excerpts, 1 and 3 are more critical than 2. For excerpt 1 and 3 the stability increases consistently with time resolution of the coder for both MDCT and FFT. For excerpt 2 an FFT with medium time resolution shows best gradings.

The performance of the MDCTs and FFTs with the same length is virtually identical.

4.3 Quality, Disregarding Image Distortions

In task 3 the audio quality is assessed without the image degradations. The results in Fig. 9 show that excerpt 2 and 3 appear critical for certain coders. For excerpt 2 (singing) the distortions occurring with the 256-point

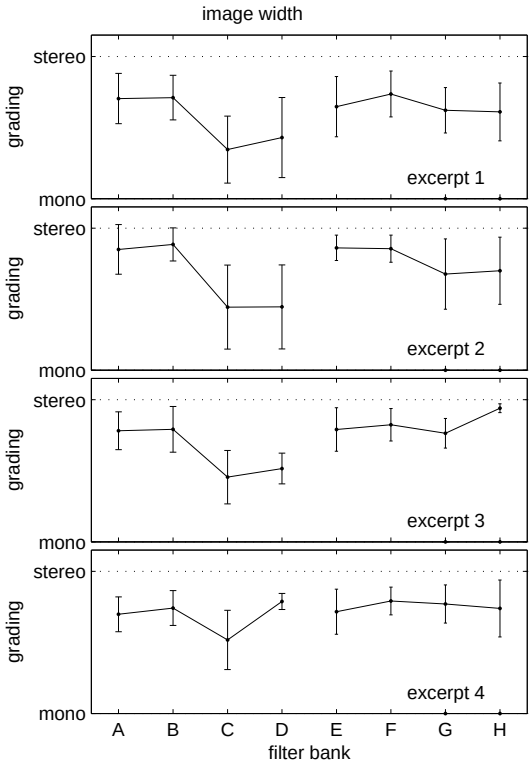

Fig. 7: Subjective gradings and 95%-confidence intervals of image width with respect to the reference (task 1).

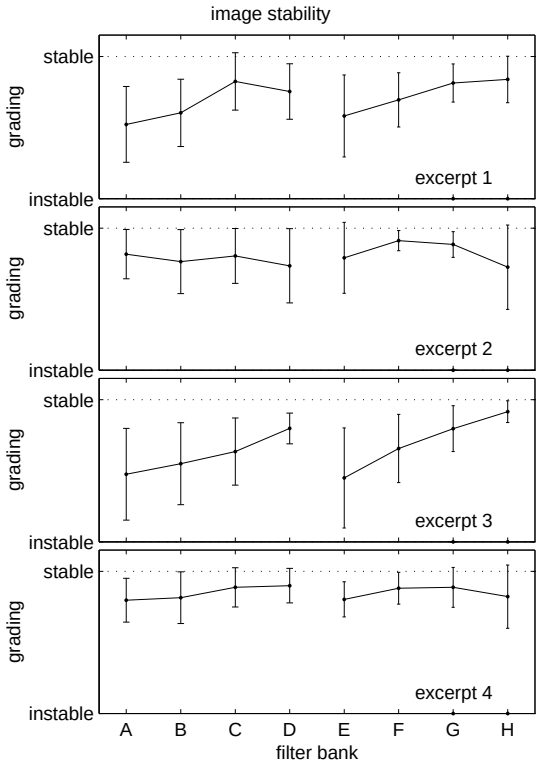

Fig. 8: Subjective gradings and 95%-confidence intervals of image stability with respect to the reference (task 2).

MDCT are identified as aliasing. Without oversampling the aliasing distortions are larger than with an oversampling factor of 2. For the FFT-based coder, aliasing distortions are not detected.

For excerpt 3 (percussions) distortions are observed for the low-time-resolution coders with 1024 or 2048 block size.

4.4 Overall Quality

The overall quality gradings in Fig. 10 show the integral impact of all noticeable distortions on audio quality to facilitate the selection of the coder with best overall performance. Obviously, the overall quality reflects the influence of the degradations assessed in task 1, 2, and 3 and it combines these individual components into a perceptually meaningful global measure.

From visual inspection it is concluded that the 256-point FFT-based coder has best performance for the test excerpts followed by the 512-point FFT. The other coders show significantly reduced quality for at least one excerpt. The 256-point FFT has a clear advantage over longer FFTs for excerpt 3 (percussions) which requires a high time resolution. For the more stationary excerpts 1 and 2, an FFT length between 256 and 1024 reaches about the same quality.

The long 2048-point MDCT has insufficient time resolution. The short 256-point MDCT performs worse than the FFT of the same length because it creates noticeable aliasing distortions.

5 SUMMARY AND CONCLUSIONS

In this paper we outlined the commonalities and differences of binaural cue coding (BCC) and Intensity Stereo Coding (ISC). While ISC is integrated into the core audio coder, BCC can independently apply a spectral signal representation that is best suited for coding the auditory spatial image. The core audio coder uses a critically sampled spectral representation using the MDCT which can create aliasing distortions if ISC is applied before the reconstruction, especially for short 256-point MDCT blocks. For long 2048-point MDCT blocks the time resolution is not high enough to preserve the spatial image

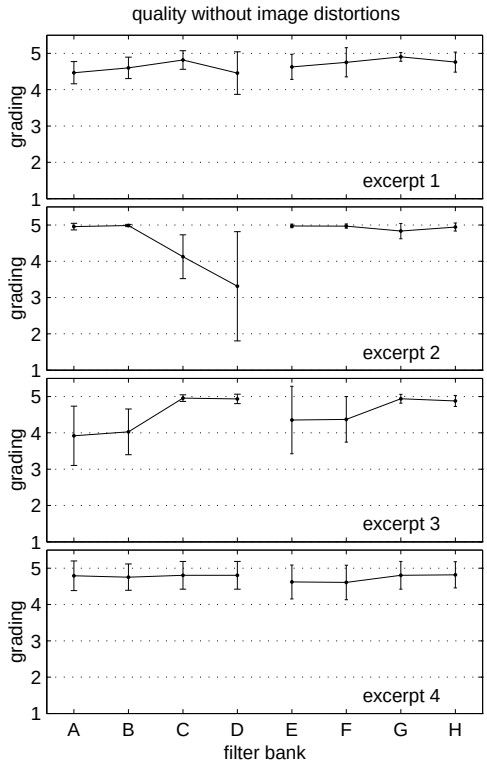

Fig. 9: Subjective gradings and 95%-confidence intervals of audio quality, disregarding image distortions with respect to the reference (task 3).

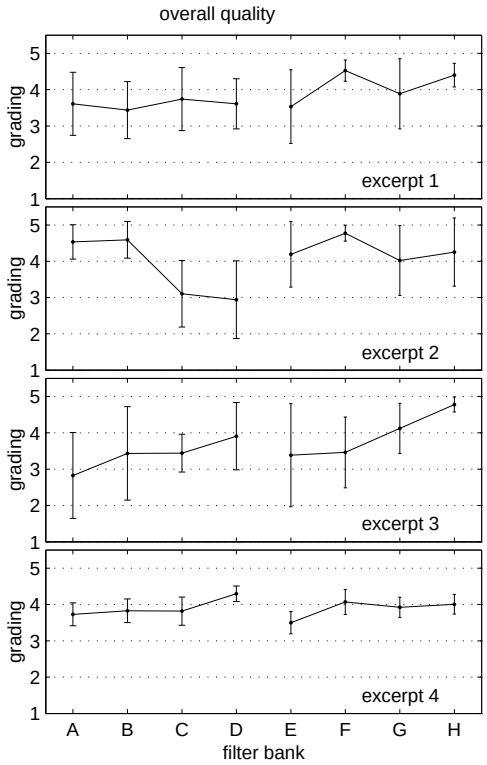

Fig. 10: Subjective gradings and 95%-confidence intervals of overall quality with respect to the reference (task 4).

stability of signals with short attacks, like percussions.

The spectral representation of the BCC implementation used here is based on the FFT [5]. Using the FFT with overlapping windows implies an oversampled sub-band representation that reduces aliasing components in comparison with the critically sampled MDCT representation. Different FFT lengths of 256, 512, 1024, and 2048 are subjectively evaluated. The best overall performance is obtained with the high time-resolution 256-point FFT. This FFT size delivers robust performance since it does not show noticeable aliasing distortions and its time resolution is high enough to suit the quickly changing spatial image of the "applause" recording and percussions.

The reported results for ISC are intentionally based on the application of ISC to the full audio bandwidth. Typically, ISC is only used above a frequency of a few kHz to avoid some of the artifacts described. For high quality coding ISC is only enabled for less critical audio segments as analyzed by the encoder. Furthermore, ISC can differ from the fixed time/frequency resolution used here, since the core coder will typically use block switching to achieve a higher time resolution, if necessary. Therefore, it might be unlikely that a percussion or applause recording will be encoded using only a 2048-point MDCT. However, the purpose of this paper is to point out the systematic performance differences of ISC and BCC and their reasons. As opposed to ISC, BCC can be applied to the full bandwidth using a static filter bank resolution. In this mode of operation, ISC performs significantly worse than BCC. Since the bit rate of BCC or ISC-encoded spectral components is lower compared with independently encoding each channel, BCC potentially leads to lower bit rates than ISC since it can be applied permanently to the full bandwidth. The side information rate per sub-band is supposed to be equal for ISC and BCC.

Since only the basic ISC and BCC systems are addressed here, there are several ways for future enhancements. Some of these enhancements include a better adapted time/frequency resolution for BCC which can be non-uniform or based on switching between resolutions. Preliminary experiments show furthermore that

aliasing artifacts can be reduced by proper smoothing of the frequency-dependent level differences without affecting the spatial image.

6 ACKNOWLEDGMENTS

We thank Peter Kroon and Jens Meyer for helpful comments on an earlier version of this paper.

REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, *Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio*, ISO/IEC 11172-3 International Standard, 1993.

[2] ISO/IEC JTC1/SC29/WG11, *Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding*, ISO/IEC 13818-7 International Standard, 1997.

[3] R.G.v.d. Waal and R.N.J. Veldhuis, "Subband coding of stereophonic digital audio signals," *Proc. IEEE ICASSP 1991*, pp. 3601–3604, 1991.

[4] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," *96th AES Conv., Feb. 1994, Amsterdam (preprint 3799)*, 1994.

[5] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, Oct. 2001, pp. 199–202.

[6] F. Baumgarte and C. Faller, "Estimation of auditory spatial cues for binaural cue coding," in *Proc. ICASSP 2002, Orlando, Florida*, May 2002.

[7] C. Faller and F. Baumgarte, "Binaural cue coding: A novel and efficient representation of spatial audio," in *Proc. ICASSP 2002, Orlando, Florida*, May 2002.

[8] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization*, MIT Press, 1983.

[9] H.S. Malvar, *Signal processing with lapped transforms*, Artech House, 1992.

[10] J. Princen, "The design of nonuniform modulated filterbanks," *IEEE Trans. Sig. Proc.*, vol. 43, no. 11, pp. 2550–2560, Nov. 1995.